Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper PUBLISHERS key to the Researchers

Research Article
Volume 6 Issue 3 - April 2018
DOI: 10.19080/BBOAJ.2018.06.555688

Biostat Biometrics Open Acc J

An Investigation of the Discriminatory Ability of the Clustering Effect of the Frailty Survival Model

Robin Van Oirbeek* and Emmanuel Lesaffre

Department of Biostatistics and Statistical Bioinformatics, KU Leuven University, Belgium

Submission: November 01, 2017; **Published:** April 25, 2018

***Corresponding author:** Robin Van Oirbeek, Department of Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium, Tel: 0032 485 920629; Email: robin.vanoirbeek@gmail.com

Abstract

A strategy is proposed to examine the importance of the clustering effect of the frailty model by means of the concordance probability. To this end, the methodology proposed in earlier work is extended to general frailty models and a new definition of the concordance probability is developed. The resulting measures allow to separate the discriminatory ability of the covariate effects and the clustering effect on a within- and between-cluster level. Using a Bayesian and a likelihood approach, point estimates and 95% credible/ confidence intervals are computed for the measures. Estimation properties and sensitivity against misspecification are checked in an extensive simulation study. As such, the likelihood estimation procedure showed difficulties in estimating unbiasedly the predictive ability of the clustering effect, while the Bayesian estimation procedure resulted in good estimation properties for all three measures. Other features are developed to complement the earlier methodology, such as an internal validation strategy as well as a procedure to calculate internally validated interval estimates in the presence of clustered data. The ensemble of the developments is illustrated on a dental data set.

Keywords : Clustered data; Concordance probability; Discrimination; Frailty model; Predictive ability

Abbreviations : IPCW: Inverse Probability Censoring Weighted; MSE: Measures The Empirical; PH: Proportional Hazards

Introduction

The frailty model is a popular model to analyze clustered survival data. For this model, the importance of clustering typically is assessed by means of the variance of the frailty distribution only: when the variance is large, clustering is important and visa versa. In this article we would like to investigate the importance of clustering in more detail. To this end, we propose to assess how clustering affects the quality of the model predictions. Despite the importance of the frailty model, few predictive measures have been suggested for the frailty model. To our knowledge, only the concordance probability has been extended to PH frailty models, resulting in 3 separate versions of the concordance probability [1]. These 3 concordance measures allow to evaluate the discriminatory ability on a within-cluster and between-cluster level separately and to measure the predictive gain of the clustering effect on top of the covariate effects. The methodology in [1] shows, however, two major shortcomings which will be tackled in this article.

The first shortcoming in [1] is that the concordance probability is estimated using Harrell's C estimation technique. The disadvantage of this estimation technique is that, in the presence of censoring, it suffers from a bias that depends on the censoring distribution. Moreover, this estimation technique is only designed for a time-unrestricted definition of the concordance probability that assumes survival models with non-crossing survival curves. In this article, we make use of a Inverse Probability Censoring Weighted (IPCW) estimator of [2] designed for the time-restricted definition concordance probability of [3]. This IPCW estimator is proven to be consistent, given that the censoring model is correctly specified. The time-restricted definition of the concordance probability has the advantage that it can be applied to any survival model. Moreover, we also introduce a new time restricted definition of the concordance probability, bearing a slightly different interpretation than the one introduced by [3]. Note that this shortcoming has also been addressed by [4], but their IPCW estimator assumes the PH frailty model as the survival model and a censoring distribution that does not depend on covariates. The second shortcoming in [1] is that their methodology is insufficiently tailored to its use in practice. As such, we provide an internal validation procedure to correct the estimated concordance probabilities for over optimism when the data set is used for both model fitting and model evaluation. For this, we adapt the internal validation procedure proposed for the Brier score for univariate survival data [5]. We also show how one can calculate an interval estimate for the overoptimistic and the internally validated concordance probability of interest.

In Section 2.1, the time-unrestricted, the time-restricted definition of [3] and the new time-restricted definition of the concordance probability are introduced for univariate survival data. In Section 2.2, it is shown how these definitions can be adapted for frailty model in general, including interval estimates. The estimation of internally validated point estimates is dealt with in Section 2.3. The estimation properties of the two time-restricted definitions of the concordance probability using a Bayesian and a likelihood approach are investigated in Section 3.1 by means of a large-scale simulation study. The effect of misspecifying the frailty distribution and the censoring model on the estimation of the concordance measures is shown in Section 3.2. All developments are applied to a dental data set in Section 3.3, followed by a discussion in Section 4.

Methods

Different concordance probabilities

In the next section, one time-unrestricted and two time-restricted definitions of the concordance probability are shown for univariate survival data, after which for each respective definition its common estimation technique is discussed.

Definitions

[6] proposed the unrestricted concordance probability *C* which is the earliest definition of the concordance probability. It equals the probability that a randomly selected subject *i*, which fails later than another randomly selected subject *j*, has a higher predicted survival time than subject *j* or,

$$C = P\left(\hat{T}\left(X_i\right) < \hat{T}\left(X_j\right) \middle| T_i < T_j\right), \quad (1)$$

With T_i, $\hat{T}\left(X_i\right)$ and X_i the true failure time, the time-independent predicted failure time (e.g. median survival time) and the covariate vector of the failure time model of subject *i*. Since the subjects of a survival study are followed up over a limited time period, the estimation of the concordance probability should be regarded as particular to the study. The above definition, however, gives the impression that it applies to the whole time range, since no time restriction is imposed on the pairs of subjects that are evaluated in the above definition. As a result, the use of (1) in a study with a limited follow-up implies that the same survival process holds beyond the last follow-up time, which is an unverifiable assumption. Moreover, since (1) evaluates the ranking of the time-independent predicted failure times $\left(\hat{T}\left(X_i\right), \hat{T}\left(X_j\right)\right)$ of a pair (i, j), the survival probability at time point $\tau, S\left(\tau \middle| X\right)$, can only be used as a predictor when its ranking is not time dependent, which does not hold for all failure time models. A restricted version of the concordance probability C^τ was therefore introduced by [3], i.e.:

$$C^\tau = P\left(\hat{T}\left(\tau \middle| X_i\right) < \hat{T}\left(\tau \middle| X_j\right) \middle| T_i < T_j \,\&\, T_i < \tau\right), \quad (2)$$

With $\hat{T}\left(\tau \middle| X_i\right)$ representing the time-dependent predicted failure time (e.g. survival probability at a certain time point τ). Measures how well subjects, who fail before time point *τ*, can be distinguished from subjects that fail later than the considered subject. Different τ values can be chosen to investigate how earlier failing subjects are differentiated from later failing subjects over time. In practice however, it might also be of interest to examine how well subjects are discriminated when their true failure times are close to each other as compared to when their true failure times differ greatly. To investigate this, we propose a new concordance measure:

$$C^\tau\left(d\right) = P\left(\hat{T}\left(\tau \middle| X_i\right) < \hat{T}\left(\tau \middle| X_j\right) \middle| T_i < T_j \,\&\, T_i < \tau \,\&\, T_j - T_i < d\right), \quad (3)$$

With *d* the maximum difference in true failure time between the two randomly selected subjects. As such, (3) measures how well subjects with a maximum difference *d* in failure time can be distinguished from each other, given that the earliest failing member of the pair fails before time point τ. Plotting $C^\tau\left(d\right)$ versus different values of *d* provides a more detailed image of the discriminatory ability of the failure time model. Note that we propose to use the survival probability at time point τ, i.e. $S\left(\tau \middle| X_i\right)$, as $\hat{T}\left(\tau \middle| X_i\right)$ for definitions (2) and (3) since it captures the cumulative risk for the event in $[0, \tau]$.

Estimation

For a sample of size *n*, consider a subject $i(i = 1 \ldots n)$ with observed failure time $t_i = \min\left(T_i, C_i\right)$ and C_i its censoring time. The censoring indicator δ_i equals 1 if $t_i = T_i$ and 0 otherwise. The unrestricted concordance probability *C* is typically estimated by Harrell's estimation technique [6] or:

$$\hat{C} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(\hat{T}\left(X_j\right) < \hat{T}\left(X_j\right) \& t_i < t_j \,\&\, \delta_i = 1\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(t_i < t_j \,\&\, \delta_i = 1\right)}, \quad (4)$$

With x_i the observed covariates of the failure time model of subject *i*. Pairs that agree with the condition in the denominator (numerator) of the upper formula are called comparable (concordant) pairs. As proven by [2], (4) is known to suffer from an overoptimistic bias that depends on the censoring distribution. In the remainder of this article, we will no longer consider this estimation technique of the concordance probability. [2] proposed the Inverse Probability Censoring Weighted (IPCW) estimation method for the C^τ definition, which corresponds to:

$$.\hat{C}^\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(\hat{T}\left(\tau \middle| x_i\right) < \hat{T}\left(\tau \middle| x_j\right) \& t_i < t_j < \tau \,\&\, \delta_i = 1\right) \hat{\omega}^{-1}\left(t_i, z_i, z_j\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(t_i < t_j \,\&\, t_i < \tau \,\&\, \delta_i = 1\right) \hat{\omega}^{-1}\left(t_i, z_i, z_j\right)}, (5)$$

with $\hat{\omega}^{-1}\left(t_i, z_i, z_j\right) = \hat{G}\left(t_i - \middle| z_i\right)\hat{G}\left(t_i - \middle| z_j\right)$ the estimated censoring weight of the pair $(i, j), t_i -$ the time point just before t_i and $G\left(t \middle| z_i\right)$ the survival probability of the censoring model and z_i the observed covariates of the censoring model of subject at time point t. Clearly, $\lim_{\tau \to +\infty} \hat{C}^\tau = \hat{C}$ and that $\hat{\omega}(\cdot)$ s equal to 1 when all subjects are uncensored.

Just as for Harrell's estimation technique of C, (5) only considers concordant pairs in the numerator and comparable pairs in the denominator, but each pair is multiplied with its estimated corresponding censoring weight $\hat{\omega}(\cdot)$. The censoring weights are obtained by means of a censoring model, which is fitted to the data by reversing the meaning of the censoring indicator: the true censoring time is censored when the event occurred $(\delta = 1)$, and uncensored when the event did not occur $(\delta = 0)$. Further, (5) estimates C^{τ} consistently under conditional independence of the censoring mechanism and the failure time model and if the censoring model is correctly specified. Hence, because the censoring weights only depend on the censoring model, estimator (5) is robust against misspecification of the failure time model [2]. The censoring weights therefore compensate for the use of just comparable pairs in the estimation of C^{τ}, counterbalancing the overoptimistic bias that would have been obtained else wise. Note that the covariates X and Z of the failure time model and the censoring model respectively are not necessarily the same.

Following [2], we have developed the following consistent IPCW estimator of $C^{\tau}(d)$:

$$\hat{C}^{\tau}(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(\hat{T}(\tau|x_i) < \hat{T}(\tau|x_j) \,\&\, t_i < t_j \,\&\, t_i < \tau \,\&\, \delta_i = 1 \,\&\, t_j - t_i < d\right)\hat{\omega}^{-1}(t_i, z_i, z_j)\hat{\phi}(t_i, x_j, z_j, d)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left(t_i < t_j \,\&\, t_i < \tau \,\&\, \delta_i = 1 \,\&\, t_j - t_i < d\right)\hat{\omega}^{-1}(t_i, z_i, z_j)\hat{\phi}(t_i, x_j, z_j, d)},$$

With

$$\hat{\phi}(t_i, x_j, z_j, d) = \frac{\hat{G}(t_i|z_j)\hat{S}(t_i|x_j) - \hat{G}(t_i|z_j)\hat{S}(t_i + d|x_j)}{\hat{G}(t_i|z_j)\hat{S}(t_i|x_j) - \hat{G}(t_i + d|z_j)\hat{S}(t_i + d|x_j)},$$

The additional weight of the pair (i, j). Clearly, $\hat{\phi}(\cdot) \le 1$ and $\lim_{d \to +\infty}\hat{\phi}(\cdot) = 1$ such that $\lim_{d \to +\infty}\hat{C}^{\tau}(d) = \hat{C}^{\tau}$. Moreover, $\hat{\phi}(\cdot)$ and $\hat{\omega}(\cdot)$ are both equal to 1 when all subjects are uncensored. In (6), $\hat{\phi}(\cdot)$ is needed because $\hat{C}^{\tau}(d)$ requires a quantification of the probability that the difference in failure time of the considered pair is lower than d in the presence of censoring. Further, since $\hat{\phi}(\cdot)$ depends on both the failure time model and the censoring model, (6) is a consistent estimator of $C^{\tau}(d)$ only when both models are correctly specified. The derivation of (6) as well as the proof of its consistency are shown in Appendix.

Application to the frailty model

The concordance probability has already been adapted to the framework of Proportional Hazards (PH) frailty models of the power variance family for the unrestricted definition (1) of the concordance probability [1]. In this section, we will extend the concordance probability to the whole class of frailty models. In the next two sections, we restrict ourselves to definition (2). For the developments related to definition (3), we refer to Section A of the online supplementary material.

Definition

The frailty model is a popular model for clustered survival data. For each subject of cluster q, a frailty term w_q is introduced to account for clustering. These frailty terms are assumed to be sampled from a frailty distribution, $f_w(w)$, which mostly is a parametric distribution such as the gamma distribution [7]. For each frailty model two different types of survival probabilities can be defined for a given time point τ, i.e. a conditional survival probability $S_C(\tau|x_i, w_q)$ and a marginal survival probability $S_M(\tau|x_i) = \int_0^{\infty} S_C(\tau|x_i, w) f_w(w)\,dw$ [7]. Note that $S_C(\tau|x_i, w_q)$ takes covariate and cluster information into account, while $S_M(\tau|x_i)$ just accounts for covariate information. For the PH frailty model for instance, it holds that $S_C(\tau|x_i, w_q) = \wedge_0(\tau)\exp(-w_q \exp(\beta^T))$ with $\wedge_0(\tau)$ the cumulative baseline hazard at time point τ and β a vector of regression coefficients. For PH frailty models of the power variance family $S_M(\tau|X) = \exp\left(v\left|(\theta(1-v))\left[1 - \left(1 + \left(\theta \wedge_0(\tau)\exp(\beta^{\tau}X)\right)/v\right)^{1-v}\right]\right.\right)$ with Clearly $\mu, \theta > 0, 0 < v \le 1$. $S_M(\tau|X)$ depends on the choice of the frailty distribution, while $S_C(\tau|X, W)$ does not.

Two types of pairs can be identified for clustered data: an inter-cluster pair, whose members belong to two different clusters, and an intra-cluster pair, whose members belong to the same cluster. Therefore, depending on which type of pair and survival probability is used, 6 different versions of the concordance probability can be introduced for frailty models (Table 1). This results in a between marginal and a between conditional concordance probability C_{BM}^{τ} and C_{BC}^{τ}, respectively, in a within marginal and a within conditional concordance probability C_{WM}^{τ} and C_{WC}^{τ}, respectively, and in an overall marginal and an overall conditional concordance probability C_{OM}^{τ} and C_{OC}^{τ}, respectively.

Table 1: Definition of the 6 different types of the concordance probability. 'X' marks which elements are used by the considered measure, while '-' marks which elements are not.

Type Concordance Probability	Inter-cluster pair	Intra-cluster pair	$S_c(t|X,W)$	$S_M(t|X)$
Between Conditional	X	-		
Between Marginal	X	-		
Within Conditional	-	X		
Within Marginal	-	X		
Overall Conditional	X	X		
Overall Marginal	X	X		

Interpretation and interrelations

For the two overall concordance probabilities, it can be shown that $C_{OC}^{\tau} = \pi_B C_{BC}^{\tau} + \pi_W C_{WC}^{\tau}$ and $C_{OM}^{\tau} = \pi_B C_{BM}^{\tau} + \pi_W C_{WM}^{\tau}$ with $\pi_B\left(\pi_W\right)$ the probability that a pair is an inter-(intra-)cluster pair [1]. As a result, C_{OC}^{τ} and C_{OM}^{τ} do not add new information to the other 4 versions of the concordance probability. Moreover, their interpretation depends strongly on the sampling scheme of the data structure such that its value is hard to compare across studies with different clustering designs [1]. Since the properties of $C_{OC}^{\tau}\left(C_{OM}^{\tau}\right)$ depend on the properties of C_{BC}^{τ} and $C_{WC}^{\tau}\left(C_{BC}^{\tau}\text{ and } C_{WM}^{\tau}\right)$, we will not consider the overall measures C_{OC}^{τ} and C_{OM}^{τ} any further. For PH frailty models of the power variance family, it has been shown that $C_{WC}^{\tau} = C_{WM}^{\tau} \equiv C_{W}^{\tau}$ holds [1]. In Appendix B, we proof the latter result holds for each frailty model. This means that we can either use the conditional or the marginal survival probabilities in the calculation of C_{W}^{τ}. In addition, when the failure time model is correctly specified, it can be shown that $C_{BC}^{\tau} \geq C_{BM}^{\tau}$ for all frailty models [1].

Summing up, only 3 unique versions of the concordance probability can be defined for the frailty model and each of these definition focuses on a specific aspect of the frailty model. C_{BC}^{τ} measures the inter-cluster predictive ability of the covariate effects and the clustering effect, while C_{BM}^{τ} measures the inter-cluster predictive ability of the covariate effects only. Comparing C_{BC}^{τ} and C_{BM}^{τ} indicates how much C_{BM}^{τ} would increase if the clustering effect were completely captured by the covariate effects. C_{W}^{τ} measures the intra-cluster predictive ability of the cluster varying covariate effects.

Computation of point estimates

Consider a sample of size n, consisting of $J\left(\mathrm{j}=1\ldots J\right)$ clusters of size $n_j\left(i=1\ldots n_j\right)$. C_{BC}^{τ} C_{BM}^{τ} and C_{W}^{τ} are estimated as:

$$\hat{C}_{BC}^{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(\hat{S}_C\left(\tau\middle|x_{ji},\hat{w}_j\right) < \hat{S}_C\left(\tau\middle|x_{qp},\hat{w}_q\right) \&\ t_{ji} < t_{qp} < \tau \&\ \delta_{ji} = 1 \& j \neq q\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(t_{ji},t_{qp} \& t_{\ ji} < \tau \& \delta_{ji} = 1 \& j \neq q\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)},$$

$$\hat{C}_{BM}^{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(\hat{S}_M\left(\tau\middle|x_{ji}\right) < \hat{S}_M\left(\tau\middle|x_{qp}\right) \&\ t_{ji} < t_{qp} < t_{ji} \& \tau \&\ \delta_{ji} = 1 \& j \neq q\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(t_{ji},t_{qp} \& t_{\ ji} < \tau \& \delta_{ji} = 1 \& j \neq q\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)},$$

$$\hat{C}_{w}^{\tau} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(\hat{S}_{ji}\left(\cdot\right) < \hat{S}_{jp}\left(\cdot\right) \&\ t_{ji} < t_{qp} \& t_{ji} < \tau \&\ \delta_{ji} = 1\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q} I\left(t_{ji} < t_{jp} \& t_{ji} < \tau \& \delta_{ji} = 1\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)},$$

With $t_{ji}, x_{ji}, z_{ji}\delta_{ji}$ the observed failure time, the covariate vector of the failure time model and the censoring model and the censoring indicator for subject of cluster $j.\hat{S}_{ji}\left(\cdot\right)$ equals either $\hat{S}_C\left(\tau\middle|x_{ji},\hat{w}_j\right)$ or $\hat{S}_M\left(\tau\middle|x_{ji}\right)$.

Following the definition of C_{BM}^{τ} and (7), the marginal survival probability needs to be calculated for each subject at time point τ in the estimation of C_{BM}^{τ}. This survival probability is, however, hard to compute for most frailty models due to the integration step. Also, the marginal survival probability depends on the choice of the frailty distribution, while the conditional survival probability does not. In **Appendix B**, we prove that the ranking of $S_M\left(\tau\middle|X\right)$ in the calculation of C_{BM}^{τ} is the same as the ranking of $S_C\left(\tau\middle|X,W\right)$, treating the considered pair (incorrectly) as an intra-cluster pair. As a result, the evaluation of C_{BM}^{τ} can also be completed by means of the conditional survival curves only. Since C_{W}^{τ} can be obtained by determining the ranking of the conditional survival curves only, the same ranking rule for the predicted failure times can be established for C_{BM}^{τ} and C_{W}^{τ} for a class of frailty models, applying this ranking rule to inter-cluster pairs for C_{BM}^{τ} and to intra-cluster pairs for C_{W}^{τ}. As such, only the ranking of the linear predictors is needed in the estimation C_{BM}^{τ} of C_{BM}^{τ} and of the PH frailty model, since:

$$S_C\left(\tau\middle|x\ ,w_q\right) \geq S_C\left(\tau\middle|x_j,w_q\right) \Leftrightarrow \wedge_0\left(\tau\right)w_q\exp\left(\beta^T x_i\right) \leq \wedge_0\left(\tau\right)w_q\exp\left(\beta^T x_j\right) \Leftrightarrow \beta^T x_i \leq \beta^T x_j.$$

In summary, a ranking rule for the predicted failure times needs to be established for C_{BC}^{τ} using $S_C\left(\tau\middle|X,W\right)$ assuming that both members of the pair belong to a different cluster and for C_{BM}^{τ} and C_{W}^{τ} using $S_C\left(\tau\middle|X,W\right)$ assuming that both members of the pair belong to the same cluster.

Computation of interval estimates

A credible/confidence interval is constructed by means of a Bayesian/likelihood procedure, combined with the percentile non-parametric bootstrap of [8]. This bootstrapping technique has been adapted to the clustered data setting, i.e. resampling by cluster and always selecting the same number of clusters in each bootstrap sample [9]. When the failure time model is fitted using likelihood or a Bayesian approach, the censoring model needs to be fitted using the same approach. Even if credible intervals can be readily obtained for the Bayesian procedure from the posterior distribution of the respective concordance probabilities, the bootstrap technique is necessary to ensure that the resulting 95% credible interval captures the uncertainty of the estimation of the model parameters for future values of the observed covariates [1]. The description of both procedures using the BC_a method [8], which is a more refined version of the percentile non-parametric bootstrap method, can be found in Section B of the online supplementary material. Since the developments for the three measures are the same, we denote the measure of interest as C^{τ}. The Bayesian procedure then corresponds to:

a. After removing the burn-in part of the Markov chain, let the MCMC procedure run for an extra m iterations. At iteration p with $p = 1,\ldots,m$ of the MCMC sampling process determine the posterior estimate for each model parameter. For the PH frailty model, posterior estimates for the parameters of the baseline hazard function $\wedge_0(t)$, the covariate effects β and the frailty terms w are obtained in this manner [7]. This step needs to be applied to the failure time model and the censoring model separately.

b. For each subject obtain the posterior sample of the censoring weight $\omega_i(\cdot)$ based on the posterior sample of the censoring model parameters.

c. Compute $C^{\tau(l)}$ at each iteration l with $l=1,\dots,m$ based on the posterior sample of the failure time model parameters and the posterior sample of the censoring weights $\omega(\cdot)$. We suggest to take the median of the posterior values of $C^{\tau(l)}$ is the point estimate $\hat{C}_\tau$.

d. Apply the Bayesian bootstrap technique of [10] to sample $k=1,\dots,k$ new bootstrapping data sets. Calculate the point estimate of each k^{th} bootstrap data set $\hat{C}^{\tau(k)}$ based on the posterior samples $\hat{C}^{\tau(l,k)}$.

e. Compute the 95% credible interval (CI) by means of the percentile method of [8] using the point estimates of the K bootstrap data sets $\hat{C}^{\tau(k)}$. As such, the 2.5 and the 97.5 percentile of the bootstrap point estimates $\hat{C}^{\tau(k)}$ constitute the lower and the upper bound of the 95% credible interval $\hat{C}^\tau$. of

The likelihood procedure is equal to:

a) Determine the point estimates of the model parameters, which are obtained for the PH frailty model by maximizing the marginal likelihood for the parameters of the baseline hazard function $\wedge_0(t)$ and the covariate effects β after which empirical Bayes estimates are calculated for the frailty terms w [7]. This step needs to be applied to the failure time model and the censoring model separately.

b) For each subject obtain the point estimate of censoring weight $\omega_i(\cdot)$ based on the estimates of the censoring model parameters.

c) Compute the point estimate $\hat{C}^\tau$ based on the estimates of the failure time model parameters and point estimates of the censoring weights $\omega(\cdot)$.

d) Apply the bootstrap technique of [8] to sample $k=1,\dots,k$ new bootstrap data sets. Calculate the point estimates of each k^{th} bootstrap data set $\hat{C}^{\tau(k)}$.

e) Compute the 95% confidence interval (CI) by means of the percentile method of [8] using the point estimates of the K bootstrap data sets $\hat{C}^{\tau(k)}$. As such, the 2.5 and the 97.5 percentile of the bootstrap point estimates $\hat{C}^{\tau(k)}$ constitute the lower and the upper bound of the 95% confidence interval of C^τ.

Internal validation

In practice, the same data is often used for developing the failure time model and measuring the predictive ability, hereby generating possibly overoptimistic estimates of the predictive ability. One can correct for this by applying an adaptation of the .632 and .632+ bootstrap cross validation procedures of the Brier score for univariate failure time models to the clustered setting, proposed by [5]. Let $\hat{C}^\tau$ be the (overoptimistic) point estimate of the measure, the procedure then amounts to:

Additional developments are presented here for C^τ_{BC} ,C^τ_{BM} and C^τ_wSince the developments for the three measures are the same, we denote the measure of interest as C^τ.

a. Split the sample randomly in training and a validation set. Make sure that the training (validation) set contains approximatively 63.2% (36.8%) of the subjects of the original sample.

b. Repeat the same modeling steps in the training sample as in the original data set. Obtain estimates for the model parameters.

c. Calculate a point estimate of the measure $\hat{C}^\tau_{\text{boot}}$ on the subjects of the validation set only.

d. Repeat the first three steps of this procedure B times with B reasonably large.

Note that step a) of the upper procedure is different from the step a) of the classical procedure [5, 11] to avoid technical difficulties regarding the estimation of the frailty terms. However, the properties of both procedures are the same since the training set is supported on 63.2% of the original data points for both procedures. The internally validated estimate corresponds to $\left(1-\psi^\tau\right)\hat{C}^\tau+\left(\psi^\tau/B\right)\sum_{b=1}^{B}\hat{C}\tau_{\text{boot},b'}$ where $\psi^\tau=0.632$ for the .632 bootstrap cross validation estimate $\hat{C}^\tau_{.362}$ and $\psi^\tau=0.632/\left(1-.368\hat{R}^\tau\right)$ for the .632+ bootstrap cross validation estimate $\hat{C}^\tau_{.362+}$ [5]. The relative over fitting rate $\hat{R}^\tau$ is:

$$\hat{R}^\tau=\frac{\hat{C}^\tau-\dfrac{1}{B}\sum\limits_{b=1}^{B}\hat{C}^\tau_{\text{boot},b}}{\hat{C}^\tau-\hat{C}^\tau_{\text{NoInf}}},$$

Where, $\hat{C}^\tau_{\text{NoInf}}$ corresponds to the no-information error rate of the restricted concordance probability for which the survival status is independent from the covariates x. As such, for each subject i of cluster j, its contribution to is computed by averaging over all the observed values of the data set or:

$$\hat{C}^\tau_{\text{NoInf}}=\frac{\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q}\left[\sum_{k=1}^{J}\sum_{l=1}^{n_q}I\left(\hat{T}\left(\tau\middle|x_{kl}\right)<\hat{T}\left(\tau\middle|x_{qp}\right)\&t_{ji}<t_{qp}\&t_{ji}<\tau\&\delta_{ji}=1\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)\right]}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{q=1}^{J}\sum_{p=1}^{n_q}\left(t_{ji}<t_{qp}\&\tau\&\delta_{ji}=1\right)\hat{\omega}^{-1}\left(t_{ji},z_{ji},z_{qp}\right)}$$

When $\hat{R}^\tau$ is lower than 0 (higher than 1), $\hat{R}^\tau$ is fixed to 0 (1). The .632+ bootstrap cross validation estimate will differ more from the .632 bootstrap cross validation estimate as R^τ gets closer to 1. Further, [11] have found that the .632+ bootstrap

How to cite this article: Robin Van O, Emmanuel L. An Investigation of the Discriminatory Ability of the Clustering Effect of the Frailty Survival Model. Biostat Biometrics Open Acc J. 2018; 6(3): 555688. DOI: 10.19080/BBOJ.2018.06.555688.

cross validation estimate performs generally better than the .632 bootstrap cross validation estimate. An estimate of the frailty term of each of the clusters of the data set is needed to estimate C_{BC}^{τ}. However, it is possible that some clusters of the original data set are not present in the training set of step a) of the above procedure such that in the validation set of step c) no point estimates can be calculated for these subjects. Therefore, we propose to calculate C_{BC}^{τ} in step c) only for those subjects whose cluster was present in the training set of step a). Note that it is assumed that C_{BC}^{τ} cannot be calculated for only a minority of the clusters of the validation data set such that the 63.2/36.8 proportion (training/validation) is not too strongly disrupted.

A too strong disruption of this proportion invalidates the upper internal validation scheme. A credible/confidence interval for the internally validated point estimate can be constructed by combining the above procedure and the one of Section 2.2.4. More specifically, we start by sampling K bootstrap data sets. For each of these bootstrap data sets, the upper procedure is repeated resulting in K validated measures $\hat{C}_{.632}^{\tau(k)}$ and $\hat{C}_{.632+}^{\tau(k)}$ with $k=1,\ldots,k.$ These validated measures $\hat{C}_{.632}^{\tau(k)}$ and $\hat{C}_{.632+}^{\tau(k)}$ are then used to construct a 95% credible/confidence interval by applying step e) of the procedure of Section 2.2.4.

Results

Simulation study

After describing the design of the simulation study, the properties of the Bayesian/likelihood point estimates and the percentile credible/confidence intervals of C_{BC}^{τ}, C_{BM}^{τ} and C_{W}^{τ} of the parametric failure time model are presented. In Section C of the online supplementary material, a full description of additional investigations is provided. Below we only report the conclusions of the additional investigations. These additional investigations entail the effect of changing the intra-cluster correlation, Changing τ and increasing the sample size on the Bayesian/likelihood point estimates of C_{BC}^{τ}, C_{BM}^{τ} and C_{W}^{τ}. Also, the properties of the BC_{a} credible/confidence intervals of the parametric failure time model and the properties of the Bayesian/likelihood point estimates of the semi-parametric failure time model of all measures are shown in Section C of the online supplementary material as well as the fitting procedures of the Bayesian/likelihood parametric and semi-parametric failure time model. This simulation study is repeated for the $C^{\tau}(d)$ measures detailed in Section D of the online supplementary material. Below we only report the conclusions of the latter simulation study.

Design of the simulation study

In this section we provide details on the simulation study. A gamma frailty proportional hazard (PH) model with a Weibull baseline hazard (shape $\rho=2$, scale $\lambda=2.236$) was chosen as the failure time distribution. Two subject-specific covariates were considered, each covariate was sampled from a standard univariate normal distribution. The β parameters were taken as (−2.45, 2.4), representing a population with strong covariate effects. All clusters are equal in size with cluster sizes 2, 5, 10 and 20. The variance parameter of the gamma frailty distribution equals 1 representing a population with a relatively large unobserved heterogeneity. This results in 4 different populations and for each of these populations censoring times are generated from a uniform distribution $U(0,v)$ with chosen to obtain a censoring percentage of 0%, 25%, 50% or 80%. This leads to 16 scenarios for which the performance of C_{BC}^{τ}, C_{BM}^{τ} and C_{W}^{τ} (with τ equal to 1) is evaluated. For each of the scenarios, 100 (500) data sets of size 1000 were necessary for the estimation of the bias and MSE (coverage probability) to obtain stable results. All population parameters are chosen such that the covariates have an appreciable effect on the failure time distribution.

In order to evaluate the properties of C_{BC}^{τ}, C_{BM}^{τ} and C_{W}^{τ}, the true values $C_{BC}^{\tau,T}$, $C_{BM}^{\tau,T}$ and $C_{W}^{\tau,T}$ are computed empirically in the absence of censoring by sampling a large data set (N = 20,000) from each of the 4 populations and by calculating the C_{BC}^{τ}, C_{BM}^{τ} and C_{W}^{τ} using the true values of the parameters of the failure time. The fitted failure time model of the parametric model corresponds to a PH model with a Weibull baseline hazards. The fitted failure time model of the semi-parametric model corresponds to a classical Cox PH model with a partial likelihood estimation procedure [12] for the likelihood procedure and to a gamma independent increments model for the baseline hazards and a PH specification for the covariates [13] for the Bayesian procedure. Since no covariates influence the censoring process, the Kaplan-Meier model of [14] is chosen as the censoring model for the likelihood procedure and the gamma-independent increments model of [13] for the Bayesian procedure. The empirical bias is calculated as the averaged difference between the true and the estimated concordance probabilities and the MSE as the mean of the squared differences between the true and estimated concordance probabilities. A positive (negative) empirical bias is defined as any estimate greater (smaller) than the true value of $C_{BC}^{\tau,T}$, $C_{BM}^{\tau,T}$ and $C_{W}^{\tau,T}$ respectively.

Simulation results for the parametric and semi-parametric model

Bayesian procedure: For the parametric model, no substantial empirical bias is seen over the different settings and that for all measures. The empirical MSE of $C_{BC}^{\tau}\left(C_{W}^{\tau}\right)$ increases mildly as the cluster size increases (decreases), attaining considerably higher levels for a censoring percentage of 80% and a cluster size of 2. For all measures the empirical MSE increases slightly as the censoring percentage increases (Figure 1). For C_{BC}^{τ} and C_{BM}^{τ}, the estimated 95% coverage probability approaches the nominal level for all scenarios. For cluster sizes 2 and 5 (10 and 20), C_{W}^{τ} attains an estimated coverage probability of about 98% (95%) (Table 2). For C_{BC}^{τ} (C_{BM}^{τ} and C_{W}^{τ}), the size of the credible interval increases as the cluster size increases (decreases) and/or the censoring percentage increases (Table 3). For the semi-parametric model, C_{BC}^{τ} shows good estimation properties for the scenarios of a cluster size of 20 only. C_{BM}^{τ} and C_{W}^{τ} are virtually unaffected by the censoring percentage and the cluster size.

Table 2: Simulation study: coverage probability of the credible/confidence intervals of C_{BC}^{τ} ,C_{BM}^{τ} and C_{W}^{τ} using the percentile method estimated by the Bayesian and likelihood method, investigated for varying degrees of censoring ('0%', '25%', '50%' and '80%') and cluster size ('2', '5', '10' or '20').

	C_{BC}^{τ}				C_{BM}^{τ}				C_{W}^{τ}			
	0%	**25%**	**50%**	**80%**	**0%**	**25%**	**50%**	**80%**	**0%**	**25%**	**50%**	**80%**
Bayesian												
2	97.2	96.2	96.2	96.8	96.4	97.6	97.6	97	98.4	97.6	97.6	98.6
5	95.2	94	96.2	95.2	97	97.2	96.4	97	97.8	98.4	97.8	98.4
10	94.8	93.8	95.4	95.8	96.2	96	94.6	95.6	96.4	96.2	95.8	96.2
20	96.0	95.0	94.8	95.8	95.2	94.6	94.8	95.8	96	96.6	96.2	95.8
Likelihood												
2	89.4	87.6	89.6	88	96.8	97.4	97.8	97.2	98	98.8	98.4	97.6
5	96.2	96.6	95.8	95.4	96.4	95.6	98.6	97.6	97.8	98.2	98.2	98.2
10	94.8	95.6	94.8	95.4	96.6	97	95.4	97.2	97.2	97.4	97.6	98.2
20	95.8	95.8	94.2	96.2	97.2	95.6	95.2	95.6	95.4	96.2	97	95.4

Table 3: Simulation study: width of the credible/confidence intervals of C_{BC}^{τ} ,C_{BM}^{τ} and C_{W}^{τ} using the percentile method estimated by the Bayesian and likelihood method, investigated for varying degrees of censoring ('0%', '25%', '50%' and '80%') and cluster size ('2', '5', '10' or '20').

	C_{BC}^{τ}				C_{BM}^{τ}				C_{W}^{τ}			
	0%	**25%**	**50%**	**80%**	**0%**	**25%**	**50%**	**80%**	**0%**	**25%**	**50%**	**80%**
Bayesian												
2	0.101	0.104	0.114	0.167	0.145	0.147	0.149	0.171	0.287	0.299	0.328	0.412
5	0.111	0.113	0.113	0.138	0.137	0.151	0.153	0.167	0.184	0.197	0.192	0.248
10	0.116	0.119	0.116	0.131	0.14	0.139	0.136	0.162	0.153	0.154	0.165	0.196
20	96	95	94.8	95.8	95.2	94.6	94.8	95.8	96	96.6	96.2	95.8
Likelihood												
2	0.113	0.113	0.126	0.183	0.147	0.158	0.156	0.179	0.305	0.317	0.333	0.431
5	0.121	0.121	0.123	0.14	0.153	0.153	0.161	0.179	0.206	0.201	0.212	0.266
10	0.13	0.124	0.127	0.144	0.148	0.149	0.152	0.169	0.166	0.167	0.174	0.21
20	0.132	0.134	0.134	0.149	0.139	0.141	0.142	0.164	0.141	0.148	0.157	0.183

How to cite this article: Robin Van O, Emmanuel L. An Investigation of the Discriminatory Ability of the Clustering Effect of the Frailty Survival Model. Biostat Biometrics Open Acc J. 2018; 6(3): 555688. DOI: 10.19080/BBOJ.2018.06.555688.

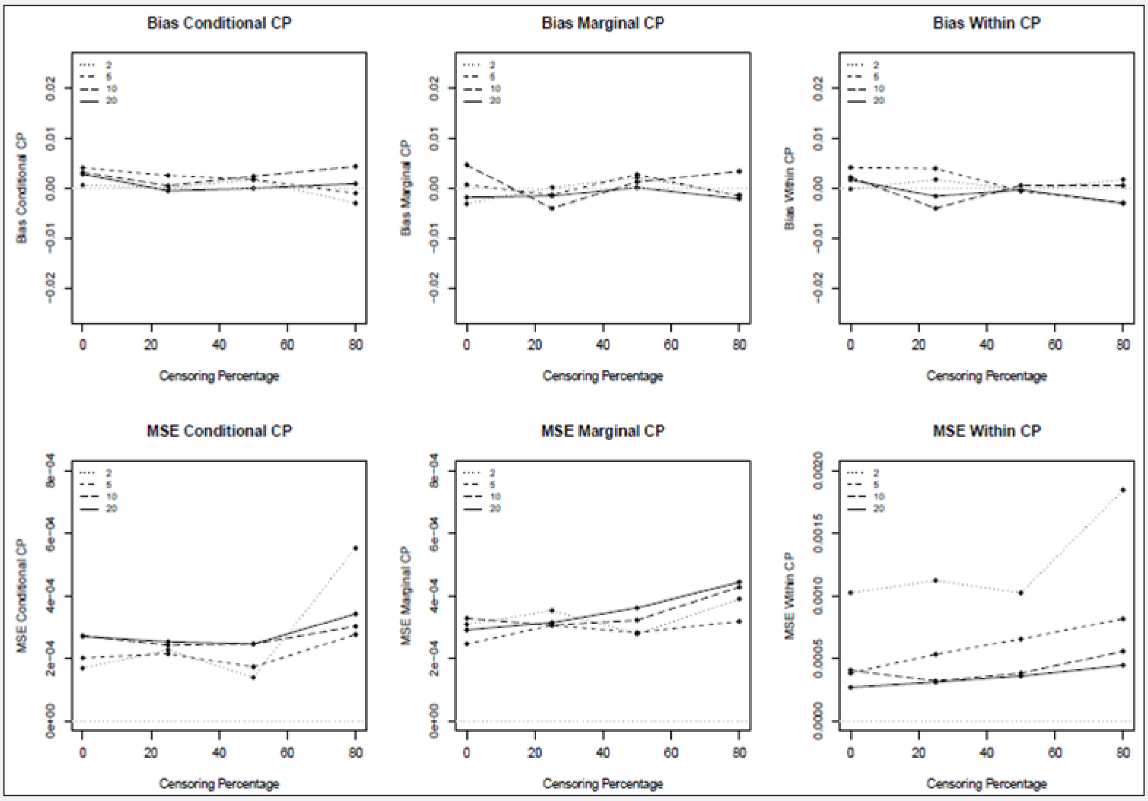

Figure 1: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating C^{τ}_{BC}, C^{τ}_{BM}, and C^{τ}_{W}, by the Bayesian method. Each plot represents a set of scenarios sharing the same cluster size (‘2’,‘5’,‘10’ or ‘20’). The X-axis refers to the censoring percentage (‘0’,‘25’,‘50’ and ‘80’) and the Y-axis to the empirical bias (upper panels) or to the empirical MSE (lower panels).

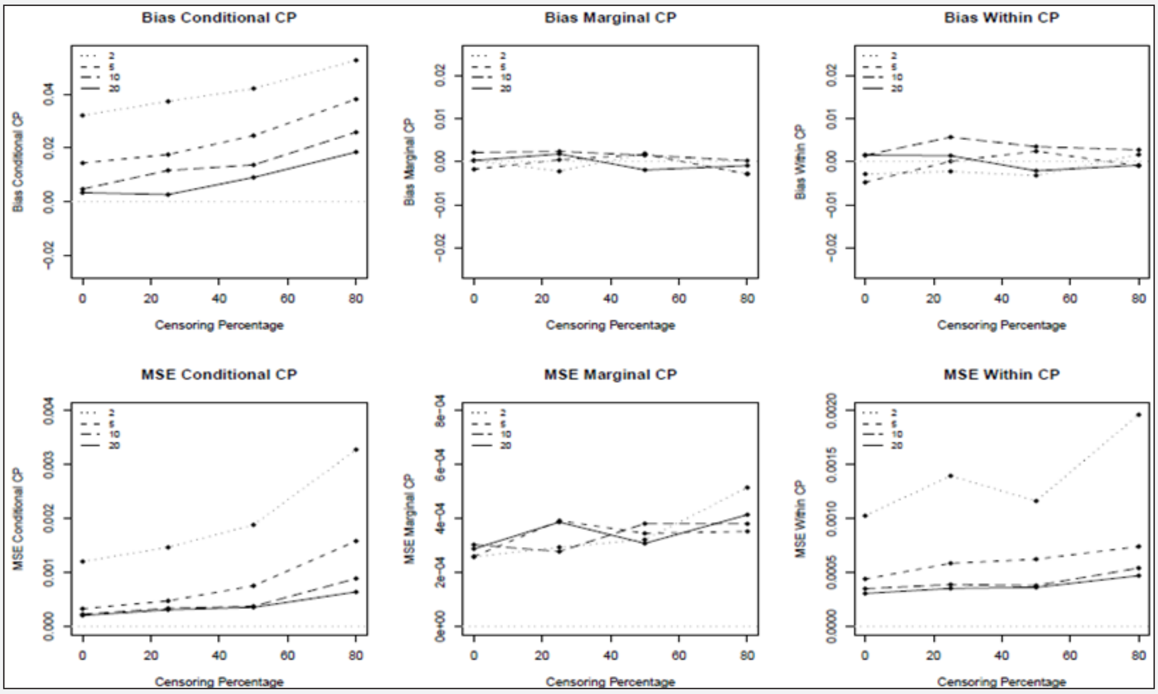

Figure 2: Simulation study: empirical bias (upper panels) or empirical MSE (lower panels) in estimating C^{τ}_{BC}, C^{τ}_{BM}, and C^{τ}_{W}, by the likelihood method.

Likelihood procedure: For the parametric model, C_{BC}^{τ} shows a positive empirical bias which increases as the cluster size decreases and the censoring percentage increases. For C_{BM}^{τ} and C_{W}^{τ} no substantial empirical bias is seen over the different settings. The empirical MSE of C_{BC}^{τ} and C_{W}^{τ} increases as the cluster size decreases, and for all measures the empirical MSE increases as the censoring percentage increases, attaining significantly higher levels for a censoring percentage of 80% (Figure 2). For C_{BC}^{τ} and C_{BM}^{τ}, the estimated 95% coverage probability approaches the nominal level for all scenarios. For cluster sizes 2 and 5 (10 and 20), C_{W}^{τ} attains an estimated coverage probability of about 98% (95%) (Table 2). For C_{BC}^{τ} (C_{BM}^{τ}, and C_{W}^{τ}), the size of the confidence interval increases as the cluster size increases (decreases) and/or the censoring percentage increases (Table 3). The semi-parametric likelihood approach shows a similar behavior as the parametric likelihood approach.

Additional simulation results for the parametric model

The estimation of C_{BC}^{τ} by the Bayesian method shows no significant empirical bias, except when the intra-cluster correlation is very small, the cluster size very small and the censoring percentage very high. The latter empirical bias is provoked by an overestimation of the frailty variance parameter. For the likelihood method, C_{BC}^{τ} shows a positive empirical bias which increases as τ and the intra-cluster correlation increases, but which is unaffected by the sample size. The estimation properties of C_{BM}^{τ} and C_{W}^{τ} by the Bayesian and likelihood method are not influenced by τ, the intra-cluster correlation or the sample size. The quality of the credible/confidence intervals of the percentile and BC_a the bootstrap method are found to be very similar. Note that the estimate of the acceleration $\hat{a}^{\tau}$ of the BC_a non-parametric bootstrap method was found to be very small in every investigated scenarios, such that we recommend to fix the $\hat{a}^{\tau}$ to zero.

Simulation results for the $C^{\tau}(d)$ measures

For the Bayesian method, similar properties for all $C^{\tau}(d)$ measures are found as for their corresponding C^{τ} measures. For the likelihood method however, an additional (negative or positive) empirical bias is observed for $C_{W}^{\tau}(d)$ and $C_{BM}^{\tau}(d)$ that increases as the censoring percentage increases. Note that this additional empirical bias is caused by the $\phi(\cdot)$ weights since it requires the estimation of the conditional survival probability of the true failure time model. Indeed, the latter estimates are constructed by means of the empirical Bayes estimates of the frailty terms which will induce in some scenarios a positive empirical bias and in other scenarios a negative empirical bias. The estimation of $C^{\tau}(d)$measures by means of the likelihood and Bayesian semi-parametric failure time model is only reliable for a sufficiently large cluster size (for a cluster of size 20 in our simulation study) and, for the likelihood method only, when the censoring percentage is not high (up until a censoring percentage of 25% in our simulation study).

Effect of misspecification

In this section, we show the effect of misspecifying the frailty distribution and the censoring model on the estimation properties of C_{BC}^{τ}, C_{BM}^{τ}, and C_{W}^{τ}. the conclusion of these investigations is reported below, a full description of the simulation study and corresponding results can be found in Section E of the online supplementary material. Moreover, just the results of the Bayesian method are reported here since similar results are found for the likelihood method. For the latter method only an additional positive bias for C_{BC}^{τ} is observed which increases as the cluster size and/or censoring percentage and/or degree of misspecification increases. This simulation study is also repeated for the $C^{\tau}(d)$ measures. For a full description of the latter simulation study and its results, we refer to Section F of the online supplementary material.

Misspecification of the frailty distribution

C_{BM}^{τ} and C_{W}^{τ} are virtually unaffected by misspecification of the frailty distribution. C_{BC}^{τ}, however, is in general strongly affected by misspecification, resulting in a positive (negative) bias when the empirical distribution of the estimated frailty terms $\hat{w}$ is more (less) variable than the distribution of the true frailty terms w. Similar results were found by [15].

Misspecification of the censoring model

A functionally misspecified covariate effect or a misspecified censoring model will result in a negative bias, especially for high censoring percentages (50% and 80% in our simulation study). Covariates that have no impact on the censoring process will not influence the quality of the estimation when included in the censoring model. Omitting important variables, however, does result in a negative bias. In practice, several equally performing censoring models can be proposed. This is however not a problem since the censoring model is merely used as a working model. Note however it is easier to detect misspecifications of the censoring model as the censoring percentage increases since, in contrast to the failure time model, a higher censoring percentage results in an increase of information for the censoring model. Similar results were found by [16].

Simulation results for the $C^{\tau}(d)$ measures

Due to the misspecification of the frailty distribution, the estimated frailty terms differ from the true frailty terms, hereby not only disrupting the bias patterns of $C_{BC}^{\tau}(d)$, but also of $C_{BM}^{\tau}(d)$ and $C_{W}^{\tau}(d)$ since the estimated frailty terms are also used in the calculation of the $\phi(\cdot)$ weights. We therefore recommend to use $C^{\tau}(d)$ the measures only to appreciate a good fitting model in more detail. The results of the misspecification of the censoring model are very similar to what was obtained for the C^{τ} measures.

How to cite this article: Robin Van O, Emmanuel L. An Investigation of the Discriminatory Ability of the Clustering Effect of the Frailty Survival Model. Biostat Biometrics Open Acc J. 2018; 6(3): 555688. DOI: 10.19080/BBOJ.2018.06.555688.

Application to the amalgam data set

In this data set, the effect of different treatment modalities on the longevity of different types of amalgam restorations is investigated. The primary covariates of the study are 4 cavity wall treatments (‘CSA’, ‘CWF’, ‘Copalite’ and ‘Silver Suspension’) and the alloy of the amalgam (levels: ‘NTD’, ‘Tytin’ and ‘CNG’). The secundary covariates are the type of the restoration (levels: ‘MO/DO’ and ‘MOD’), the type of tooth (levels: ‘premolar’ and ‘molar’), the position of the tooth in the mouth (levels: ‘upper left’, ‘upper right’, ‘lower left’ and ‘lower right’), the operator (levels: ‘operator 1’, ‘operator 2’ and ‘operator 3’), gender and age. The data set is composed of three clinical trials and in each clinical trial the primary variables were combined into 4 treatment modalities assigning each modality randomly to 4 (or 8 or 12) teeth in a 2 x 2 (factorial) design within patient [17]. Note that this study was mainly designed to evaluate significance of the primary covariates. In this paper we will evaluate the predictive potential of all the primary covariates. Since the early failures are suspected to be of a different nature than the late failures and since there are only a few early failures, we restrict our analysis to the late failures only, i.e. failures that occur at least 9 years after enrollment in the study. This results in 174 patients contributing with 1347 amalgam restorations. The clustered data structure is unbalanced, with 4, 8 and 12 restorations seen in 36, 89 and 33 patients, respectively. The median follow-up time is 14.48 and the true failure time is observed for 187 amalgam restorations only, leading to a censoring percentage of 86.2%.

Based on the results of the simulation studies of Section 3.1, we have chosen to use the Bayesian gamma frailty generalized gamma model as the failure time model, which is a very flexible parametric Bayesian model [18]. Note that the failure time model consists of the primary covariates only. Different censoring models for the IPCW weights were tested with the Bayesian generalized gamma model of both primary and secondary covariates resulting into the highest concordance probability estimates. We evaluate the model’s discriminatory ability from 9 to 15 years after enrollment in the study, since, as recommended by [19], censoring is not too heavy at the end of this time interval. The +.632 bootstrap cross validation estimators are chosen for the internal validation of the considered concordance probabilities. The R code of this analysis but applied to fictive data can be found in the online supplementary material.

Analysis using the full information

The apparent and validated point estimate and 95% percentile credible interval of the C^{τ} measures (B = 100, K = 100, $\tau = 6$) are shown in Table 4. Since τ is chosen to be 6, all comparable pairs of the [9,15] time interval are considered. Little difference with the values in Table 4 is found for $C^{\tau}(d)$ measures using $\tau = 6$ and d=6. We see that the covariates have practically no intra-cluster predictive ability but a strong inter-cluster predictive ability. Further, even after validation, the between conditional concordance probability improves reasonably strongly on the between marginal concordance probability implying the presence of a strong clustering effect upon the covariate effects. Thus, if one succeeds in a future study to fully capture the clustering effect in covariates and if the primary variables are also included in this future model, the between marginal concordance probability can potentially attain a value of 0.74. Note that the drop in predictive ability due to the internal validation is more pronounced for the between conditional concordance probability than for the between marginal and within concordance probability. All the apparent and validated 95% percentile credible intervals show a strong overlap and the upper bound of the validated 95% percentile credible intervals are all lower than the upper bound of the apparent 95% percentile credible intervals. Surprisingly, the lower bound of the validated 95% percentile credible intervals of the within and the between marginal measures are higher than the upper bound of the apparent 95% percentile credible intervals.

Table 4: Amalgam data set: Apparent and validated point estimate (‘Point’) and 95% percentile credible interval (‘CI’) of C^{τ}_{BC} C^{τ}_{BM} , and C^{τ}_{W} .

	C^{τ}_{BC}	C^{τ}_{BM}	C^{τ}_{W}
	Apparent		
Point	0.801	0.685	0.506
CI	[0.736,0.928]	[0.477,0.807]	[0.260,0.706]
	Validated		
Point	0.743	0.674	0.484
CI	[0.639,0.865]	[0.534,0.798]	[0.337,0.635]

Analyzing τ and $d-$ varying patterns

In Figure 3, the apparent and validated time-varying C^{τ} and $C^{\tau}(d)$ estimates are shown. In the left panel of Figure 3, no C^{τ} estimates are available for a τ value lower than 10.2. Indeed, C^{τ} can only be estimated when at least one true failure time has been observed prior to τ, which is only the case from 10.2 years onwards. Further, C^{τ}_{BC} (C^{τ}_{BM} and C^{τ}_{W}) attain lower (higher) values as τ increases, implying that early failing teeth are easier (more difficult) to discriminate from later failing teeth than later failing teeth are. As a result, the difference between the validated curves of C^{τ}_{BC} and C^{τ}_{BM} mildly decreases as τ increases such that the clustering effect induces a stronger gain in predictive ability for the earlier failing teeth as compared to the later failing teeth. Note that C^{τ}_{W} even attains values lower than 0.5 for τ values lower than 14, meaning that the ranking of the predictions are systematically in the wrong direction for teeth belonging to this τ range. In the right panel of Figure 3, a monotone increasing pattern is depicted for both $C^{\tau}_{BC}(d)$ and $C^{\tau}_{BM}(d)$. The validated patterns of both measures exceed 0.5 only for d values higher than 0.7 and remain to be very similar up until a d value of 1. Stable results are obtained for $C^{\tau}_{BC}(d)$ and $C^{\tau}_{BM}(d)$ once d exceeds 3 years. This means that the clustering effect only increases upon the predictive ability of the covariates for inter-cluster pairs with a difference in failure time of minimum 1 year and of maximum 3 years. The validated pattern of $C_W^{\tau}(d)$ approaches a straight line and reveals an almost complete absence of intra-cluster predictive ability of the primary covariates for this data set.

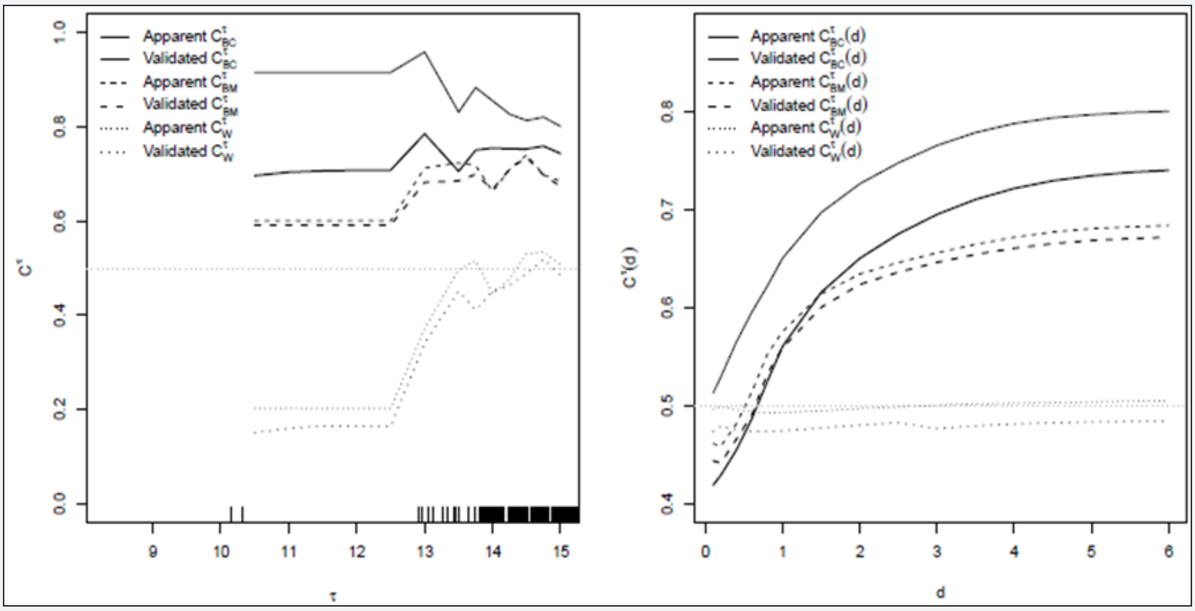

Figure 3 : In the left (right) panel, the apparent and validated $\tau(d)$ varying patterns of the $C\left(C^{\tau}(d)\right)$ estimates are shown. Ticks for the observed true failure times are shown in the left panel.

Discussion

In this article, the methodology proposed in [1] is extended to general frailty models. Also, a new definition of the concordance probability is developed as well as an internal validation procedure and a procedure to calculate interval estimates in the presence of clustered data. Note that all these developments can also be applied to univariate survival models, we just focused on the frailty model in this article. Other very useful measures can be extended to the frailty model in a similar manner [19-22]. We chose the concordance probability since it is most widely used measure for the discriminatory ability of survival data. From the simulation study, we learned that it is of mayor importance to model the censoring model correctly. Despite of this possible shortcoming, the IPCW estimation method is up to now, to our knowledge, the best method to estimate the concordance probability in the presence of censoring. An alternative estimation method could be based on a non-parametric imputation method based on the covariate distribution, in which the censored failure times are replaced by true failure times whose covariate information is close to the one of the censored observation. This estimation would be, in spirit, similar to what has been done by [23] for the estimation of survival models.

The between conditional concordance probability estimated by the likelihood approach suffers from a negative bias, especially for a small cluster size and a high censoring percentage. By means of a small simulation study, we could determine that this overoptimistic bias is caused by the empirical Bayes estimates of the frailty terms since the overoptimistic bias of the between conditional concordance probability disappeared once the frailty terms were replaced by their population values. Note that similar results were found in [1] and [24]. An external validation of the between conditional concordance probability can only be performed when the same clusters are used for the external data set as for the training data set. Indeed, only under these conditions estimates of the frailty terms for the clusters of the external data set are available, indispensable to compute the between conditional concordance probability of the external data set. Note that this only emphasizes the importance of an internal validation procedure for the between conditional concordance probability. A different solution to this problem could be the estimation of the frailty term of a new cluster, but this approach was not considered here. In this article we extended the concordance probability for frailty models with a 2-level clustering structure, since this is the most common clustering structure in survival analysis [7]. In practice however, more complicated clustering structures may be encountered. In [24] for example, it is shown how the concordance probability and the Brier score can be extended for nested 2-level and 3-level multilevel binary regression models.

Acknowledgement

The authors would like to thank Cees Kreulen for the permission to use the amalgam data set [17] for this research. For the computations, we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules

Foundation and the Flemish Government - department EWI. Support from the IAP project Phase VII (P07/6, 2012-2017) to develop crucial statistical methods for understanding major complex dynamic systems in natural, biomedical and social sciences from the Federal Office for Scientific, Technical and Cultural Affairs of Belgium is gratefully acknowledged.

References

1. Van Oirbeek R, Lesaffre E (2010) An application of Harrell's C-index to PH frailty models. Stat Med 29(30): 3160-3171.
2. Gerds TA, Kattan MW, Schumacher M, Changhong Y (2013) Estimating a time dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med 32(13): 2173-2184.
3. Heagerty PJ, Zheng Y (2005) Survival model predictive accuracy and ROC curves. Biometrics 61(1): 92-105.
4. Maugain A, Collette S, Pignon JP, Rondeau V (2013) Concordance measures in shared frailty models: application to clustered data in cancer prognosis. Stat Med 32(27): 4803-4820.
5. Gerds TA, Schumacher M (2007) Efron-type measures of prediction error for survival analysis. Biometrics 63(4): 1283-1287.
6. Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. J AMA 247(18): 2543-2546.
7. Duchateau L, Janssen P (2008) The Frailty Model. Springer Science + Business Media, New York, USA.
8. Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman & Hall/CRC: London, Boca Raton, FL, USA.
9. Field CA, Welsh AH (2007) Bootstrapping clustered data. JR Statist Soc B 69(part 3): 369-390.
10. Rubin DB (1981) The Bayesian bootstrap. The Ann Statist 9(1): 130-134.
11. Efron B, Tibshirani RJ (1997) Improvements on cross-validation: the .632+ bootstrap method. Journal of the American Statistical Association 92(438): 548-560.
12. Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society, Series B (Methodological) 34(2): 187-220.
13. Kalbfleisch JD (1978) Nonparametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society, Series B 40(2): 214-221.
14. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53(282): 457-481.
15. Glidden DV, Vittinghoff E (2004) Modelling clustered survival data from multicentre clinical trials. Stat Med 23(3): 369-388.
16. Gerds TA, Schumacher M (2006) Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom J 48(6): 1029-1040.
17. Kreulen CM, Tobi H, Gruythuysen RJM, van Amerongen WE, Borgmeijer PJ (1998) Replacement risk of amalgam treatment modalities: 15-year results. J Dent 26(8): 627-632.
18. Cox C, Chu H, Schneider MF, Muñoz A (2007) Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine 26(23): 4352-4374.
19. Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18(17-18): 2529-2545.
20. Schemper M, Stare J (1996) Explained variation in survival analysis. Stat in Med 15(19): 1999-2012.
21. O Quigley J, Xu R, Stare J (2005) Explained randomness in proportional hazards models. Stat in Med 24(3): 479-489.
22. Pencina M, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS (2008) Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 27(2): 157-172.
23. Hsu CH, Taylor JMG, Murray S, Commenges D (2006) Survival analysis using auxiliary variables via non-parametric multiple imputation. Stat Med 25(20): 3503-3517.
24. Van Oirbeek R, Lesaffre E (2012) Assessing the predictive ability of a multilevel binary regression model. Computational Statistics & Data Analysis 56(6): 1966-1980.